Measuring teaching skills in elementary education using the Rasch model

van de Grift, Wim J. C. M.; Houtveen, Thoni A. M.; van den Hurk, Henk T. G.; Terpstra, Oscar

Published in:
School Effectiveness and School Improvement

DOI:
10.1080/09243453.2019.1577743

Publication date:
2019

School Effectiveness and School Improvement
An International Journal of Research, Policy and Practice

ISSN: 0924-3453 (Print) 1744-5124 (Online) Journal homepage: https://www.tandfonline.com/loi/nses20

Measuring teaching skills in elementary education using the Rasch model

Wim J. C. M. van de Grift, Thoni A. M. Houtveen, Henk T. G. van den Hurk & Oscar Terpstra

To cite this article: Wim J. C. M. van de Grift, Thoni A. M. Houtveen, Henk T. G. van den Hurk & Oscar Terpstra (2019) Measuring teaching skills in elementary education using the Rasch model, School Effectiveness and School Improvement, 30:4, 455-486, DOI: 10.1080/09243453.2019.1577743

To link to this article: https://doi.org/10.1080/09243453.2019.1577743

Published online: 10 Apr 2019.

SCHOOL EFFECTIVENESS AND SCHOOL IMPROVEMENT
2019, VOL. 30, NO. 4, 455–486
https://doi.org/10.1080/09243453.2019.1577743

Routledge
Taylor & Francis Group

ARTICLE

OPEN ACCESS Check for updates

Measuring teaching skills in elementary education using the Rasch model

Wim J. C. M. van de Grift [a], Thoni A. M. Houtveen [b], Henk T. G. van den Hurk [b] and Oscar Terpstra [b]

[a]Department of Teacher Education, University of Groningen, Groningen, The Netherlands; [b]Research Centre for Education, University of Applied Sciences, Utrecht, The Netherlands

ABSTRACT

Observation scales for measuring teaching skills were developed for both elementary education and kindergarten. Based on 500 observations, we found that both scales meet the requirements of the dichotomous Rasch model. These observation scales can help in finding the zone of proximal development of teachers in elementary education and kindergarten. This can help in improving teachers' skills.

ARTICLE HISTORY
Received 6 February 2018
Accepted 30 January 2019

KEYWORDS
Effective teaching; elementary education; kindergarten; observation; Rasch model; zone of proximal development

Introduction

Half a century ago, teacher evaluation was mostly regarded by teacher training institutes as a tool to help decide whether a student teacher was properly prepared for the job. Later, teacher evaluation acquired a more central place in various international policy documents and reports (e.g., Commissie Evaluatie Basisonderwijs [Committee Evaluation Primary Education], 1994; Department for Education, 2012; Doherty & Jacobs, 2013; Mourshed, Chijioke, & Barber, 2010).

The use of teacher evaluation methods received more emphasis in policy documents, after research showed that about 15% to 25% of the difference in student achievement can be ascribed to the work of teachers (Aaronson, Barrow, & Sander, 2007; Brandsma & Knuver, 1989; Houtveen & Van de Grift, 2007a, 2007b; Houtveen, Van de Grift, & Brokamp, 2014; Houtveen, Van de Grift, & Creemers, 2004; Rockoff, 2004; Roeleveld, 2003; Wijnstra, Ouwens, & Béguin, 2003). As a result, a wide range of research-based classroom observation instruments has since been developed.

Nowadays, teacher evaluation serves three aims. These are no longer restricted to policy aims, but now include formative and summative aims. Summative teacher evaluation supports decisions about teacher selection and decisions about the progress of a teacher's career. However, it is often "forgotten" that reliable summative decisions require more than 10 observations done by different observers (Van der Lans, Van de Grift, Van Veen, & Fokkens-Bruinsma, 2016). Formative evaluation also requires various observations from different observers in order to arrive at reliable decisions. In a coaching situation, this

CONTACT Wim J. C. M. van de Grift W.J.C.M.van.de.Grift@rug.nl Department of Teacher Education, University of Groningen, Groningen, The Netherlands
This article has been republished with minor changes. These changes do not impact the academic content of the article.

problem is usually solved by having a short conversation with the observed teacher, asking questions like: "Was the lesson representative?" or "Did the teacher have the opportunity to show all of his skills?" If the answer to these questions is "no", then a second observation or a second opinion is required.

This study focusses on the development of an observation instrument that is designed to be used in a formative (coaching) context, though it may also be applied in a summative context or for the purposes of educational research.

For use in a formative context, it is important that the observation tool offers possibilities to make a detailed assessment of those skills that are just out of reach for the teacher. The point is that the observation instrument should reveal the skills that the teacher **just** does not show, but that are not too difficult for him to learn. With a nod to Vygotskij (1934/2002), we propose saying that a good observation instrument should reveal a teacher's zone of proximal development. In this article, we will present an observation instrument that might be helpful for the incremental coaching of teachers in their zone of proximal development.

Theoretical and empirical background

Observation instruments

Since the 1960s, many observation instruments have been developed to evaluate the quality of teaching (Capie, Johnson, Anderson, Ellet, & Okey, 1980; Evertson, 1987; Evertson & Burry, 1989; Flanders, 1961, 1970; Florida Coalition for the Development of a Performance Measurement System, 1983; Houtveen, Booij, De Jong, & Van de Grift, 1999; Houtveen & Overmars, 1996; Inspectie van het Onderwijs, 1998; Office for Standards in Education, 1995; Schaffer & Nesselrodt, 1992; Slavin, 1987; Stallings, 1980; Stallings & Kaskowitz, 1974; Teddlie, Virgilio, & Oescher, 1990; Tricket & Moos, 1974; Veenman, Lem, Voeten, Winkelmolen, & Lassche, 1986; Virgilio, 1987; Virgilio & Teddlie, 1989). Most of these instruments were originally developed for teacher training rather than research purposes. Nowadays, we have various research instruments that meet the high demands of reliability and validity. Some important observation instruments are presented below.

The Classroom Assessment Scoring System (CLASS) was developed by Pianta, LaParo, and Hamre (2008) at the University of Virginia to assess classroom quality in PK-12 (preschool, kindergarten to 12th grade) classrooms. CLASS was originally designed for early-year contexts, but was later extended to include the full age range in education. CLASS describes multiple dimensions of teaching that are linked to student achievement and development. CLASS has three broad domains: emotional support, classroom organization, and instructional support. The CLASS observation instrument can be used to assess classroom quality for both research purposes and as a tool to help new and experienced teachers to become more effective.

The Framework for Teaching (FfT) was developed by Danielson in 2007 and revised in 2013 (Danielson, 2011). The Framework for Teaching is a research-based set of components of instruction, grounded in a constructivist view of learning and teaching. The complex activity of teaching is divided into 22 components (and 76 smaller elements) clustered into four domains of teaching responsibility: planning and preparation, classroom environment, instruction, and professional responsibilities.

The International Comparative Analysis of Learning and Teaching (ICALT) observation instrument was originally developed by Van de Grift and colleagues for use in the national inspection system in the Netherlands (Van de Grift, 1985, 2007; Van de Grift & Lam, 1998). The instrument includes 32 high-inferential and 120 low-inferential items that specify observable teaching behaviours, grouped in six domains: safe learning climate, classroom management, clear instruction, activating teaching methods, learning strategies, and differentiation.

The International System for Teacher Observation and Feedback (ISTOF) was developed by a large team of international researchers, based on the accumulated knowledge from several decades of research in the educational effectiveness tradition (Teddlie, Creemers, Kyriakides, Muijs, & Yu, 2006). The ISTOF framework contains 11 components of quality, of which seven are assessed through classroom observation: assessment and evaluation, differentiation and inclusion, clarity of instruction, instructional skills, promoting active learning and developing metacognitive skills, classroom climate, and classroom management.

The Mathematical Quality of Instruction (MQI) is an observational rubric specific to mathematics, constructed by Heather Hill and colleagues at the University of Michigan and Harvard University. It is used to measure several dimensions of teaching students mathematical content (Hill et al., 2008; Hill, Umland, Litke, & Kapitula, 2012). The MQI is based on a theory of instruction, existing literature on effective instruction in mathematics, and an analysis of the diverse teaching methods of hundreds of teachers in the United States. MQI has five domains: common core-aligned student practices, working with students and mathematics, richness of mathematics, errors and imprecision, and classroom work related to mathematics.

The Generic Dimensions of Teaching Quality model was developed by Klieme and colleagues in Germany, within the context of the Trends in International Mathematics and Science Study (TIMSS) Video (Klieme, Schümer, & Knoll, 2001). It has been used in 18 research studies in Germany, Switzerland, and Austria, 12 of which were based on high-inference observation protocols, 6 on student and/or teacher questionnaires. It has also been adapted on an international scale for use in the projects of the Organisation for Economic Cooperation and Development (OECD, 2012, 2014), such as the Teaching and Learning International Survey (TALIS) and the Programme for International Student Assessment (PISA), as well as for school inspection in Germany, most prominently in Hamburg. It has three basic dimensions of quality: structure, clarity, and classroom management; challenge and cognitive activation; supportive climate.

Some of these observation instruments were originally designed to be used in a specific context, like young children or mathematics. Other observation instruments are more generic in character.

The ICALT observation tool has several advantages over other instruments:

- The instrument is generic in character: It can be reliably used in classes with young and older students, in lessons in a variety of subject matters, and in schools in different countries (Van de Grift, 2007, 2014),
- The high- and low-inferential items of the instrument are based on the results of research into the effects of learning and teaching (Cotton, 1995; Creemers, 1991, 1994; Ellis & Worthington, 1994; Levine & Lezotte, 1990, 1995; Muijs & Reynolds, 2010; Purkey & Smith, 1983, 1985; Sammons, Hillman, & Mortimore, 1995; Scheerens, 1989, 1992, 2008; Van de Grift, 1985; Walberg & Haertel, 1992; Wright, Horn, & Sanders, 1997).

- Previous research has shown that the items have a particular sequence in terms of difficulty that is in accordance with the Rasch model (Van de Grift & Lam, 1998; Van de Grift, Van der Wal, & Torenbeek, 2011).

It is for these reasons that we have decided to develop further the ICALT observation tool, to use it for detecting the zone of proximal development of observed teachers.

Feedback

The use of feedback in the classroom has gathered a great deal of research attention over the last decade. The models applied in this research are often based on the feedback model developed by Hattie and Timperley (2007). In this model, feedback is defined as the information given by a person (e.g., teacher, mentor, colleague, or others) with regard to aspects of a person's performance or knowledge. Feedback is therefore a consequence of the performance. Feedback is information with which a learner can confirm, expand, replace, or change the information stored in memory (Butler & Winne, 1995). The provided feedback can relate to knowledge as well as to opinions about yourself and the performance of tasks, or to behaviour. Feedback has no effect within a vacuum. To be effective, there must be a learning situation in which the feedback is given. The question is how effective feedback is. In 12 meta-analyses of 196 studies into the effectiveness of education, feedback in the classroom was included as an influencing variable (Hattie & Timperley, 2007). In these studies, the average effect size of the contribution of feedback to learning performance was .79. Feedback, in addition to direct instruction and mutual teaching, is one of the most important factors contributing to learning achievements. However, the effect sizes found in the studies in question vary widely, indicating that some forms of feedback are more powerful in comparison than others. The largest effect sizes were found in studies where students were given feedback on the task they had performed (ES: 1.10) as well as when they were given instructions on how to improve their performance of the task (ES: .94). The effect of giving compliments (ES: .14), rewarding (ES: .31), and penalties (ES: .20), on the other hand, is much smaller. The most effective forms of feedback included giving concrete instructions, and/or the feedback was clearly related to the goals to be achieved (Hattie & Timperley, 2007). Feedback is more effective when it contains information about what was right and why, and when it builds upon what went before it; in other words, when it is regularly given and related to what should be learned. Feedback also has the most impact when the goals to be achieved are specific and challenging, but the task to be performed to achieve the goal is not overly complex (Kluger & DeNisi, 1996). Against the background of the available knowledge about the effectiveness of feedback, Hattie and Timperley (2007) developed a model for feedback aimed at promoting learning. The main purpose of feedback in this model is to reduce the discrepancy between the current situation or performance and a goal. For feedback to be effective, three main questions always need to be answered: What are my goals? (Feed Up), what progress has been made in achieving the goals? (Feed Back), and what activities should be undertaken to get closer to the goal? (Feed Forward) (Hattie & Timperley, 2007, p. 86).

For our study, however, the important question is: What effects have been found regarding giving feedback to *teachers*? Houtveen (1990) has examined the effects of

counselling on teacher behaviour, and Thurlings (2012) has studied the mutual feedback from teachers. In a study conducted by Van den Hurk, Houtveen, and Van de Grift (2016), which reported the effect sizes of the growth of teaching skills after feedback on their lessons, it was found that among teachers who received feedback on their lessons, the effect size of skill growth, depending on the observed aspect, ranged from .29 (creating a safe and stimulating climate) to three quarters of a standard deviation for activating students and teaching learning strategies.

The next question is: Does the improvement of teaching skills go hand in hand with an increase in the learning gain of students? Several small-scale field experiments, with experimental and control groups and a pre- and post-test design, have been carried out in order to improve the learning gain of students in elementary education with regard to decoding, comprehensive reading, and mathematics. The treatment of these experiments consisted in teaching and training teachers in several teaching skills based on classroom observation and coaching. The teachers in the experimental conditions showed a growth in their teaching skills of one quarter to more than a full standard deviation. The learning gains (corrected for gender, age, intelligence, and socioeconomic status) of the students in these experimental groups exceeded the learning gains in the control groups; for decoding by 28% and 62% of a standard deviation, for comprehensive reading by 52% of a standard deviation, and for mathematics by 36% of a standard deviation (Houtveen & Van de Grift, 2007a, 2007b; Houtveen et al., 2004). It is therefore clear that it would be worthwhile to invest in a classroom observation instrument that is suitable for giving feedback to teachers in their zone of proximal development.

Research aim

The aim of this study is to develop an observation instrument that might be helpful for the incremental coaching of teachers. More precisely, we will be investigating whether there is a specific order in the 32 items of the ICALT instrument that may be helpful in detecting the zone of proximal development of teachers in elementary education.

Method

Short history of the ICALT observation instrument

Between 1985 and 1994, various original studies aimed at the effectiveness of teachers' didactic actions were reviewed (Creemers, 1991; Levine & Lezotte, 1990; Purkey & Smith, 1983, 1985; Scheerens, 1989, 1992; Van de Grift, 1985; Walberg & Haertel, 1992). From the original studies mentioned in these reviews, all teaching activities were "sieved" that were related to the performance and/or learning gain of students, and an observation instrument was constructed based on these activities. In the 1990s, this observation instrument was used by the Dutch Education Inspectorate to evaluate the quality of Dutch primary education (Commissie Evaluatie Basisonderwijs, 1994; Van de Grift, 1994). This was the first version of what later became the "ICALT observation instrument".

In the years that followed, more original studies were conducted that focussed on the effectiveness of teachers' didactic behaviour. Many of these studies were included in reviews providing an overview of the state of knowledge about the effectiveness of learning and

teaching (Cotton, 1995; Creemers, 1994; Ellis & Worthington, 1994; Levine & Lezotte, 1995; Muijs & Reynolds, 2010; Sammons et al.,1995; Scheerens, 2008; Wright et al., 1997). These new reviews, and especially the original studies into the effectiveness of teachers' didactic behaviour, led to 152 different activities all related to learning achievements and/or student gains, all of which are suitable for conducting observations in the classroom. These activities were reformulated as items and used to construct the ICALT instrument. A version of this instrument was then developed for observing teachers in primary education (Van de Grift, 2007, 2014; Van de Grift & Lam, 1998; Van de Grift et al., 2011). Later versions were also developed for beginning teachers in secondary education (Van de Grift, Helms-Lorenz, & Maulana, 2014) and experienced teachers in secondary education (Van der Lans, Van de Grift, & Van Veen, 2018; Van der Lans et al., 2016).

In this article, we will present a version of the ICALT instrument that is useable for elementary education and that fulfils the demands of the dichotomous Rasch model. The ICALT observation instrument is presented in Appendix 1.

Sample characteristics

All teachers in the sample (N = 500) had attained a teaching certification from a University of Applied Sciences (PABO), which is a form of higher professional education at a bachelor level. The teachers participating in this study were also all enrolled in a master course of a full year (60 ECTS). Table 1 offers some descriptive background statistics of the teachers in the sample.

These 500 teachers were distributed among 412 schools. In 2016, the Dutch population of schools for elementary education consisted of 6,347 schools. Our sample of schools is therefore 6.5% of the population. Table 2 presents an overview of some of the regional characteristics of the schools in our sample, in comparison with the schools in the population. The sample shows some underrepresentation of the schools in the north of the Netherlands.

Procedure

Prior to the start of the first module of the master course, all teachers were assigned to make a video or a digital recording of one of their lessons. During the first lesson of the module, all recorded lessons were independently observed by two trained peer teachers. Immediately after the observations, the observers discussed the observational data in order to reach a consensus regarding the scores. The agreed-upon scores were then entered in a digital version of the observation instrument. This digital interface ensured the recording of the results and, at the same time, provided feedback reports. In these feedback reports, the scores on the observation instrument were returned to both the observed teachers and their observers.

Table 1. Some characteristics of the teachers.

Percentage male teachers	9.2	
Percentage of teachers working with 4–5-year-olds (Kindergarten)	19.4	
Percentage of teachers working with 6–7-year-olds	37.3	
Percentage of teachers working with 8–12-year-olds	43.3	
Average amount of years of experience	9.57	*SD* 7.94
Average class size	21.46	*SD* 6.85

Table 2. Some characteristics of the schools.

Percentage of schools	Population	Sample
In the north of the Netherlands	17.7	5.4
In the middle of the Netherlands	60.8	69.9
In the south of the Netherlands	21.6	24.6

Training of observers

The training was based on the use of 32 high- and 120 low-inferential items. The 32 high-inferential items are the core of the observation instrument. The sum score on these 32 items is the raw score on the instrument. These 32 items have a more abstract (or high-inferential) character. An example of a high-inferential item is: "The teacher adjusts instructions to relevant inter-learner differences".

The 120 low-inferential items can also be observed during a lesson. During the observations, the actions in these low-inferential items are simply seen or not seen. For example, low-inferential items that belong to the high-inferential item above are: "The teacher puts learners who need little instruction to work", "The teacher gives additional instructions to small groups or individual learners", or "The teacher does not simply focus on the average learner". The low-inferential items are used to reach consensus on the scoring of the high-inferential items.

All participating observers attended a half-day training session, during which information on the theoretical and empirical backgrounds of the instrument was provided. During the training, the observers watched and entered their scores for two video-recorded lessons. The results of the first video were used for discussions between the observers. The goal was twofold: mutual agreement between the observers and agreement with an external standard based on the scores of experienced observers. During the training, observers with item scores that differed (significantly) from the majority of the observers or the external standard were invited to explain their score on the high-inferential items, along with their scores on the low-inferential items. In this way, observers were given the opportunity to learn to attach the same meaning to each high-inferential item. The results of the observations of the second recorded lesson were used to test whether the mutual agreement reached at least a moderate/good mutual consensus (Fleiss's κ of $> .60$) and a disagreement with a norm group of expert observers lower than an effect size (Cohen's δ) of .20. The scores of observers with strong deviations from these norms were excluded from the study.

The ICALT instrument and the rasch model

The most stringent item response model (IRT), the Rasch model (Rasch, 1960, 1961), offers unique possibilities for arranging items and persons along a single dimension. Every item and each person is given their specific place on this single dimension. This is very useful for finding the zone of proximal development of an observed teacher. The Rasch model requires the data of a scale to satisfy three assumptions:

- The scale is one-dimensional.
- The items of the scale are local stochastic independent.
- The item characteristic curves are parallel.

We used several different procedures in order to test whether our data fit the three assumptions of the Rasch model. First, we will report on the use of procedures derived from classical test theory, like explorative and confirmatory factor analysis. Furthermore, we also used procedures specially developed for testing the assumptions of the Rasch model, like the Andersen's log-likelihood ratio test and W.-H. Chen and Thissen's LD chi-square index for testing local dependence. In particular, where possible, we also used "graphical tests", like the scree plot of eigenvalues, as well as visualisations, like estimating the slope parameters of the item characteristic curves. The results of these procedures are reported in the following sections.

One-dimensionality

The assumption of one-dimensionality states that observations of the items on a scale can be ascribed to a single latent construct, in our case: teaching skill. Although the one-dimensionality assumption of a (Rasch) scale is difficult, if not impossible, to really prove, we can use several procedures to test whether it is likely that a set of items form a one-dimensional scale. In the following sections, we describe the results of the confirmatory factor analysis, perform a "graphical test" by making a scree plot of the eigenvalues based on the correlation matrix of items, and check whether variables other than the intended latent dimension – teaching skill – affect the item discrimination parameters.

Confirmatory factor analysis. We carried out a confirmatory factor analysis (CFA) with a one-factor model, using the statistical programme Mplus 7.4 (Muthén & Muthén, 1998–2015). The usual χ^2-based test is substantially affected by sample size (Marsh, Balla, & McDonald, 1988). Therefore, for our large sample of observations, we used the comparative fit index (CFI) and the Tucker-Lewis index (TLI). Both these indices are less vulnerable to sample size. Furthermore, we used the root mean square error of approximation (RMSEA) to assess model fit. The norms for acceptable fit, for both CFI and TLI, are > .90, and for RMSEA < .08 (F. Chen, Curran, Bollen, Kirby, & Paxton, 2008; Hu & Bentler, 1999; Kline, 2005; Marsh, Hau, & Wen, 2004; Tucker & Lewis, 1973). The results are presented in Table 3.

Both the CFI and the TLI are above the norm of .90 and the root mean square error of approximation (RMSEA) is below the norm of .08. This is an indication of one-dimensionality.

Table 3. Confirmatory factor analyses on 32 items and 1 factor.

	CFI	TLI	RMSEA
Model fit for one-dimensionality Norm	>.90	>.90	<.08
Result	.92	.92	.05

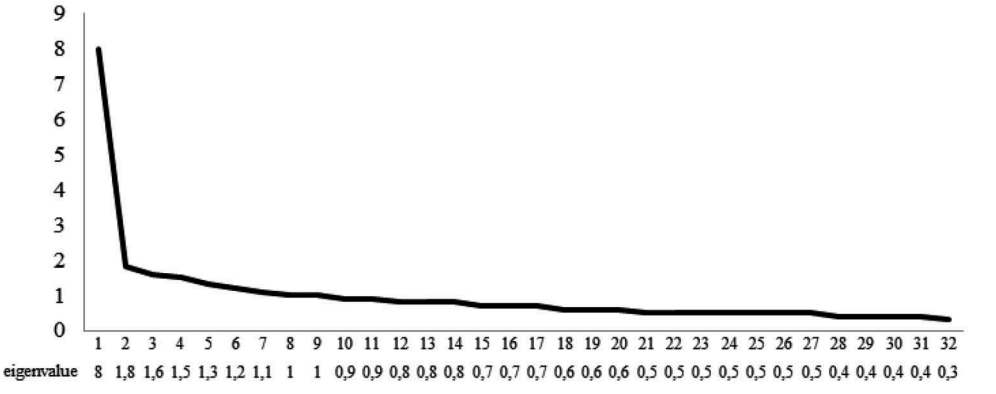

Figure 1. A scree plot of eigenvalues.

A scree plot of eigenvalues. We carried out a "graphical test" by making a scree plot of the eigenvalues based on the correlation matrix of the 32 items of the ICALT observation instrument. The eigenvalues of a factor analysis are plotted in Figure 1.

The first eigenvalue (8.0) is considerably larger than the second (1.8) and third (1.6) eigenvalues. The scree plot clearly shows one dominant factor, which is a second indication that the assumption of one-dimensionality is reasonable.

Andersen's log-likelihood ratio test. The core of a Rasch scale is that each item finds its place on only one dimension. In a one-dimensional scale, the item-difficulty parameters should not diverge for groups of teachers differing in gender, amount of teaching experience, class size, or age group of students. This is often called differential item functioning. Differential item functioning can be tested for ordinal items with a test designed by Liu and Agresti (1996), and for polytomous items with tests designed by Camilli and Congdon (1999), Mantel (1963), Penfield and Algina (2003), and Zwick, Thayer, and Mazzeo (1997). For dichotomous items that are used in the Rasch model, tests designed by Andersen (1973, 1977), Cox (1958), and Fischer and Scheiblechner (1970) can be used. We used Andersen's (1973, 1977) log-likelihood ratio test to compare the difficulty parameters (log δ) for each item. As norms, we used a p value of .01 for a moderate fit and .05 for a good fit. The Andersen test is implemented in the eRm R package (Mair & Hatzinger, 2007). The results are shown in Table 4.

The test results showed that the difficulty parameters of male and female teachers do not differ too much. The same applies to item parameters for beginning and experienced teachers and for differences in class size. However, the difficulty parameters of several items differ for teachers working in kindergarten and teachers working with other student age groups.

Conclusions about one-dimensionality. Both confirmative factor analysis and the scree plot of eigenvalues gave important indications that the assumption of one-dimensionality is reasonable. In addition, another important indication for the one-dimensionality of the scale is the fact that the item difficulty parameters did not differ, within reasonable boundaries, for male or female teachers, beginning or more experienced teachers, and teachers working in small or large classrooms. However, several

Table 4. Andersen's log-likelihood ratio test for different teacher characteristics.

	A-χ^2	*df*	*p* value
Gender	43.85	29[1]	.04
Gender leaving out Item 31	37.76	28	.10
Teaching experience (< 6 years versus ≥ 5 years of experience)	26.75	31	.69
Class size (< 21 students versus ≥ 20 students)	28.32	30[1]	.55
Kindergarten vs. Group 3–4[2]	88. 65	31	.00
Kindergarten vs. Group 3–4 leaving out Item 24	69.31	30	.00
Kindergarten vs. Group 3–4 leaving out Items 24, 32	55.79	29	.00
Kindergarten vs. Group 3–4 leaving out Items 24, 32, 25	44.02	28	.03
Kindergarten vs. Group 3–4 leaving out Items 24, 32, 25, 17	35.56	27	.09
Kindergarten vs. Group 5–8	87.88	31	.00
Kindergarten vs. Group 5–8 leaving out Item 32	75.32	30	.00
Kindergarten vs. Group 5–8 leaving out Items 32, 27	64.92	29	.00
Kindergarten vs. Group 5–8 leaving out Items 32, 27, 24	54.50	28	.00
Kindergarten vs. Group 5–8 leaving out Items 32, 27, 24, 12	49.58	27	.01
Kindergarten vs. Group 5–8 leaving out Items 32, 27, 24, 12, 17	45.03	26	.01
Kindergarten vs. Group 5–8 leaving out Items 32, 27, 24, 12, 17, 25	39.90	25	.03
Kindergarten vs. Group 5–8 leaving out Items 32, 27, 24, 12, 17, 25, 18	35.78	24	.06
Group 3–4 vs. Group 5–8	57.18	31	.00
Group 3–4 vs. Group 5–8 leaving out Item 16	44.03	30	.05
Group 3–4 vs. Group 5–8 leaving out Items 16, 27	39.56	29	.09

[1]Item 1 and/or Item 2 had no variance in one of the groups.
[2]Group 3 starts when the students are about 7 years old; in Group 8, the students are about 12 years old.

items do have different difficulty parameters for teachers working in kindergarten and teachers working with older students. Most of the infringing items are about differentiated instruction and teaching learning strategies. It is therefore necessary to leave out these infringing items with observations of teachers working in kindergarten.

Local dependence

Another of the three assumptions of the Rasch model is local stochastic independence. The assumption of local stochastic independence means that significant correlations between items disappear when the effect of the intended latent variable (in this case: teaching skill) has been partialled out. Local independence and one-dimensionality are (of course) highly interrelated. We tested the assumption of local independence with confirmatory factor analysis and with the Chen and Thissen test (W.-H. Chen & Thissen, 1997).

Confirmatory factor analysis with all residual correlations set at 0. In order to check whether the correlations between the items had disappeared after the effect of the latent skill was partialled out, we used confirmatory factor analysis. We formulated a one-factor model in which all residual correlations were set at zero. In order to test this hypothesis, we used the statistical program Mplus 7.4 (Muthén & Muthén, 1998–2015). Table 5 shows the results.

Table 5 shows that both the CFI and the TLI are above .90 and the RMSEA is below the norm of .08, which is an indication for local independence.

Table 5. Confirmatory factor analyses on 32 items and 1 factor residual correlations set at 0.

	CFI	TLI	RMSEA
Model fit for residual correlations set at 0 Norm	>.90	>.90	<.08
Result	.93	.93	.05

The LDχ2 index. W.-H. Chen and Thissen (1997) proposed a standardized index, the $LD\chi^2$ index. This index may be used to establish whether there is a violation of the assumption of local stochastic independence for item pairs. When the Chen-Thissen $LD\chi^2$ has a value > 10, it indicates possible local dependence. We computed Chen-Thissen's $LD\chi^2$ using the statistical programme IRTPRO (Cai, Thissen, & Du Toit, 2005–2013).

Results show that four pairs of items seem to be locally dependent:

(1) "clearly specifies the lesson aims at the start of the lesson" with "evaluates whether the lesson aims have been reached" ($LD\chi^2$:16.7).
(2) "stimulates the building of self-confidence in weaker learners" with "offers weaker learners extra study and instruction time" ($LD\chi^2$:22.5).
(3) "adjusts instruction to relevant inter-learner differences" with "offers weaker learners extra study and instruction time" ($LD\chi^2$:55.4).
(4) "adjusts instruction to relevant inter-learner differences" with "adjusts the processing of subject matter to relevant inter-learner differences" ($LD\chi^2$:44.6).

For all pairs, it is very clear that the mutual partial correlations have to do with a teacher's rationality. It is illogical to evaluate whether lesson aims have been reached, if lesson aims have not been specified. The same is true for the other pairs of items that have to do with differentiated teaching. The relatively high rest correlations seem to be related to the rational behaviour of teachers and not with the interference of another dimension.

Conclusions about local dependence. Confirmative factor analysis with all residual correlations set at zero produced a good model fit. This is an indication that the items in the scale are local independent. Although further analysis with Chen and Thissen's $LD\chi^2$ index showed some residual correlations in four pairs of items, these correlations do not seem to point to interference from another dimension. On the basis of the performed analyses, we decided not to leave out one or more items for reasons of local dependency.

Parallelism of item characteristic curves

The probability of a positive score on an item depends on the teaching ability of a teacher. If the probability of a positive score on an item were plotted against the skill of teachers, the result would be a smooth S-shaped curve. This is called the item characteristic curve. The item characteristic curves of the items should be parallel, indicating that the difficulty of the items remains the same for teachers with different abilities. We used various procedures to check whether this was the case for the 32 items in the scale. As a first check, we applied a procedure that has been in use for about 40 years: Andersen's log-likelihood ratio test for teachers with low and high scores. As a second check, the real slope parameters were computed with a relatively new statistical procedure.

Andersen's log-likelihood ratio test for teachers with low and high scores. Andersen's (1973, 1977) log-likelihood ratio test may be used to examine the equality of the item parameters of teachers with a high and a low skill level. With the statistical eRm R package (Mair & Hatzinger, 2007), we computed the difficulty parameters for each

Table 6. Andersen's log-likelihood ratio test for teachers with low and high scores.

	A_{icc} $\chi 2$	*df*	*p* value
32 items	50.99	31	.00
Leaving out Item 24	39.49	30	.03
Leaving out Item 24 and Item 25	31.10	29	.15

A_{icc} $\chi 2$ = Andersen's log-likelihood ratio test for testing equality of the item characteristic curves (icc).

item of both teachers with low and high scores, and compared these parameters with Andersen's log-likelihood ratio χ^2 test. The results are shown in Table 6.

These results show that with all 32 items, Andersen's log likelihood ratio χ^2 test is 50.991, with 31 degrees of freedom and a *p* value of .002, indicating some misfit. Leaving out Item 24, "offers weaker learners extra study and instruction time", shows a χ^2 that is relatively small, given the number of degrees of freedom (30). The *p* value is now .03, indicating a moderate fit. Leaving out Item 25, "adjusts instruction to relevant inter-learner differences", indicates a good fit (*p* value .15). The remaining 29 items have about the same discrimination parameters for teachers with either high or lower levels of teaching skill. This is a first indication of parallelism in these 29 item characteristic curves.

The slopes of the item characteristic curves. The Rasch model offers more simple possibilities for interpreting results with a difficulty parameter as well as with a slope parameter, than the two-parameter model (Birnbaum, 1968). However, the Birnbaum model offers the opportunity to compute the slope parameter of each item. This makes it possible to check whether the slope parameters are parallel. We used the LTM R package (Rizopoulos, 2006) to perform the Birnbaum model for estimating the slopes of the item characteristic curves. The results are shown in Table 7.

The average slope (a-parameter) is 1.73. The slopes of two items (1 and 3) are steeper (1.96 x *SE*) than the average slope parameter. Item 22 is more flat (–1.96 x *SE*) than the average slope parameter. Steep slopes at the beginning of a scale do not disturb the measurement process too much. This is the case with Items 1 and 3. However, a flat slope in the middle of the measurement scale brings along serious measurement problems. This is the case with Item 22.

Conclusions about parallelism. It is important to notice that Items 24 and 25, which were detected with Andersen's log-likelihood ratio test for teachers with low and high scores, do not show an aberrant slope parameter. In fact, the item parameters of these two items are very near to the average slope parameter. More important is the very flat slope parameter of Item 22. It became evident that it was better leave out Item 22.

Conclusions about the fit of the rasch model

First, we noticed that the slope parameter of Item 22 was too flat. This is problematic because this item has a difficulty parameter that lies more or less in the middle of the scale. This hinders the precise measurement of teaching skill, in that part of the scale where we find most of the observed teachers. It was therefore better to leave out this item.

Table 7. Slopes of the item characteristic curves.

	The teacher …	slope (a)	*SE*
1	Shows respect for learners in his/her behaviour and language	3.09	.23
2	Maintains a relaxed atmosphere	2.01	.27
3	Promotes learners' self-confidence	2.68	.26
4	Fosters mutual respect	1.85	.28
5	Ensures the lesson proceeds in an orderly manner	1.53	.25
6	Monitors to ensure learners carry out activities in the appropriate manner	1.20	.40
7	Provides effective classroom management	1.96	.28
8	Uses the time for learning efficiently	2.35	.29
9	Presents and explains the subject material in a clear manner	2.64	.63
10	Gives feedback to learners	1.59	.26
11	Engages all learners in the lesson	1.62	.39
12	During the presentation stage, checks whether learners have understood the subject material	1.49	.36
13	Encourages learners to do their best	1.82	.23
14	Teaches in a well-structured manner	1.97	.39
15	Gives a clear explanation of how to use didactic aids and how to carry out assignments	1.52	.25
16	Offers activities and work forms that stimulate learners to take an active approach	1.54	.24
17	Stimulates the building of self-confidence in weaker learners	1.89	.20
18	Stimulates learners to think about solutions	1.90	.33
19	Asks questions which stimulate learners to reflect	1.63	.32
20	Let learners think aloud	1.65	.17
21	Gives interactive instructions	1.82	.26
22	Clearly specifies the lesson aims at the start of the lesson	.96	.22
23	Evaluates whether the lesson aims have been reached	1.01	.34
24	Offers weaker learners extra study and instruction time	1.15	.36
25	Adjusts instruction to relevant inter-learner differences	1.35	.28
26	Adjusts the processing of subject matter to relevant inter-learner differences	1.48	.34
27	Teaches learners how to simplify complex problems	1.76	.24
28	Stimulates the use of control activities	1.43	.27
29	Teaches learners to check solutions	1.54	.51
30	Stimulates the application of what has been learned	1.18	.26
31	Encourages learners to think critically	1.88	.20
32	Asks learners to reflect on practical strategies	1.91	.35

Second, we noticed a different functioning of some items in the kindergarten scale in comparison with the rest of primary education. It was therefore necessary to leave out those items (12, 17, 18, 24, 25, 27, and 32) that pertained to differentiated instruction and teaching learning strategies. These items do not seem to be useful for observations of teachers working in kindergarten.

We therefore propose using a 31-item version of the ICALT instrument for use in classrooms with students of 6 years and older and a special 24-item version for use in kindergarten classrooms. This 24-item version can, of course, also be used in the classrooms with the older students, but this version will give only a modest picture of the more advanced teaching skills related to differentiated instruction and teaching learning strategies.

Results of the ICALT31 version

In this section, we describe the item difficulties and the person parameters of the ICALT31 version. Information on the shorter ICALT24 version can be found in Appendix 1.

Item difficulties and person parameters

The eRm R package (Mair & Hatzinger, 2007) was used to compute the difficulty parameters for each of the 31 selected items. We chose to estimate the person parameters using Warm's (1989) weighted likelihood estimates. Warm's procedure is less biased in comparison with the traditional maximum likelihood estimates method (Hoijtink & Boomsma, 1995). Furthermore, this procedure has the advantage that it can also be used to estimate the skills of people with a zero and a maximum score. The programme WINMIRA (von Davier, 1994) was used to compute the person parameters of Warm's weighted likelihood estimates. Table 8 shows the Wright map of the ICALT scale, with the difficulty parameters and the standard errors of the items, and the person parameters with their standard errors and relative frequency.

We have split the items into seven groups, which largely correspond to the original scales of the ICALT observation tool (cf. section on observation instruments). Small adjustments have been made to provide the groups of items with whole numbers, in a simple way so that the zone of proximal development can be made visible in an uncomplicated manner (perfect > 4, almost perfect 3–4, teaching learning strategies 2–3; differentiation 1–2; basic skills −1–1; and below the basic standards of teaching < −1).

In 5.3% of the observed lessons, the θ score is below −1. In these lessons, serious problems were found in terms of realising the very basic tasks of teaching, like creating a safe educational climate and classroom management. In other words, the atmosphere in the classroom is not relaxed, and/or the lesson does not proceed in an orderly manner, and/or the time for learning is not used efficiently, and/or the teaching is not clear or not well structured. Lessons with scores below −1 cannot really be seen as situations where students can learn. From other studies we know that beginning teachers who have problems with teaching in a well-structured manner and with ensuring that the lesson proceeds in an orderly manner tend to leave the teaching job within a few years (Maulana, Helms-Lorenz, & Van de Grift, 2015; Van de Grift & Helms-Lorenz, 2013). Presuming that no "special events" happened during the observed lesson and that other specific reasons that may explain this low score are absent, the "zone of proximal development" for these teachers is working on creating a safe educational climate and maintaining orderly classroom management.

In 11.3% of the lessons, the score is between −1 and 0. In these lessons, problems arise with basic teaching tasks, such as creating a safe educational climate, maintaining efficient classroom management, and giving clear and structured instruction.

In 22.3% of the lessons, the score lies between 0 and 1. Basic teaching tasks are shown: creating a safe educational climate, maintaining efficient classroom management, and giving clear and structured instruction. In these lessons, problems are observed with the skill of activating students. Differentiated teaching and teaching students how to learn are skills beyond reach at this point.

In 21.0% of the lessons, the score lies between 1 and 2. Basic teaching tasks are shown: creating a safe educational climate, maintaining efficient classroom management, giving clear and structured instruction, as well as activating students. Differentiated teaching seems to be the next step, while teaching students how to learn is still not a shown skill.

In 25.5% of the lessons, the score lies between 2 and 3. All basic teaching tasks are exhibited: creating a safe educational climate, maintaining efficient classroom management, and giving clear and structured instruction. The same goes for activating students

Table 8. Wright map for the ICALT31 scale (N = 400 teachers teaching 6–12-year-olds).

	Item parameters			Person parameters		Freq.
Domain	Item	Log δ	*SE*	Warm's θ	*SE*	%
Beyond the basic standards of teaching	The teacher …			−5.64	1.59	0
5.3%				−4.32	.96	0
	Shows respect for learners in his/her behaviour and language	−3.79	.46			
		−3.79				
				−3.62	.78	0
				−3.12	.68	0
	Maintains a relaxed atmosphere	−2.76	.30			
				−2.72	.62	0
				−2.37	.57	0
	Presents and explains the subject material in a clear manner	−2.21	.24			
	Teaches in a well-structured manner	−2.05	.23			
	Provides effective classroom management	−2.05	.23			
	Promotes learners' self-confidence	−2.00	.22			
				−1.79	.52	1.3
	Ensures the lesson proceeds in an orderly manner	−1.61	.20			
					.50	1.5
				−1.30	.49	.8
	Uses the time for learning efficiently	−1.08	.17			.5
				−1.07	.48	1.3
Basic teaching skills				−.85	.47	1.8
11.3%	Fosters mutual respect	−.84	.16			
				−.64	.46	2.3
	Gives a clear explanation of how to use didactic aids and how to carry out assignments	−.63	.15			
	Engages all learners in the lesson	−.60	.15			
	Encourages learners to do their best	−.51	.15			
	Monitors to ensure learners carry out activities in the appropriate manner	−.43	.15	−.43	.46	1.8
	During the presentation stage, checks whether learners have understood the subject material	−.41	.15			
				−.23	.45	3.0
	Gives feedback to learners	−.15	.14			
				−.03	.45	2.5
Activating students				.17	.45	4.3
22.3%	Offers activities and work forms that stimulate learners to take an active approach	.19	.13			
	Gives interactive instructions	.26	.13			
	Asks questions which stimulate learners to reflect	.26	.13			
				.36	.45	3.0

(*Continued*)

Table 8. (Continued).

	Item parameters			Person parameters		Freq.
Domain	Item	Log δ	*SE*	Warm's θ	*SE*	%
	Let learners think aloud	.54	.13			
				.56	.45	3.5
	Stimulates the building of self-confidence in weaker learners	.66	.13			
				.76	.45	5.0
	Stimulates learners to think about solutions	.96	.12			
				.97	.46	6.5
Differentiating teaching				1.17	.46	5.3
21.0%	Offers weaker learners extra study and instruction time	1.07	.12			
	Evaluates whether the lesson aims have been reached	1.30	.12			
				1.39	.47	5.3
	Adjusts instruction to relevant inter-learner differences	1.33	.12			
				1.61	.48	5.0
	Stimulates the application of what has been learned	1.76	.12			
	Asks learners to reflect on practical strategies	1.77	.12			
	Teaches learners how to simplify complex problems	1.80	.12			
				1.84	.49	5.5
	Adjusts the processing of subject matter to relevant inter-learner differences	1.90	.12			
Teaching learning strategies				2.09	.51	8.5
25.5%	Encourages learners to think critically	2.17	.13			
				2.35	.54	7.0
	Stimulates the use of control activities	2.46	.13			
				2.65	.58	4.3
	Teaches learners to check solutions	2.68	.14			
				2.99	.63	5.5
Almost perfect7.0%				3.42	.72	3.3
				4.01	.90	3.8
Perfect 7.8%				5.19	1.50	7.8

and differentiated teaching. The next step for teachers with a score between 2 and 3 is to develop the skill of teaching students how to learn.

Seven percent of the lessons show almost all basic and advanced tasks of teaching. Teachers with a theta score of > 3 seem to master creating a safe educational climate, maintaining efficient classroom management, giving clear and structured instruction, activating students, differentiated teaching, and teaching students how to learn. Just one or two of the 31 items need some improvement.

In 7.8% of the lessons, all 31 items of the ICALT scale are scored positive.

Some descriptive statistics

Table 9 shows the average scores on the ICALT31 scale of teachers working in Groups 3 to 4, and in Groups 5–8. (Group 3 starts when the students are about 7 years old; in Group 8, the students are about 12 years old.)

As shown in Table 9, the teaching skills of male teachers do not differ significantly from the teaching skills of female teachers. We found a small (R = .12) yet significant (p = .02) correlation between years of experience and teaching skill. Due to the small size of the subsamples, the differences between the five groups of teachers with consecutive years of experience are not significant. When inspecting the effect size differences, however, we found an effect size difference of .25 between teachers with less than 1 year of experience and teachers with 1 to 5 years of experience, and an effect size difference of .16 between teachers with 1 to 5 years of experience and teachers with 6 to 10 years of experience. Cohen (1988) proposes a criterion of δ = .20 for a small effect, and Lipsey (1990) suggests a more empirically based criterion of δ = .15 for a small effect. In order to find effect sizes of δ > .15, significant on the .05 level, with a power of .80, we would need representative subsamples of more than 270 teachers. It seems worthwhile to study the relationship between years of experience and teaching skill in more detail with a larger sample.

Table 9. Descriptives of the ICALT31 scale.

	n	μ	σ	Cohen's δ	Significance
All	400	1.62	1.67		
Gender					
Male	41	1.70	1.77		
Female	359	1.60	1.66	.06	n.s.
Years of experience (R with ICALT31 = .12; sign. = .02)					
< 1 year	39	1.11	1.34		
1–5 years	139	1.50	1.65	.25	n.s.
6–10 years	103	1.76	1.71	.16	n.s.
11–20 years	85	1.73	1.76	−.02	n.s.
> 20 years	34	1.95	1.72	.13	n.s.
Age group of students					
6–7-year-olds	185	1.40	1.54		
8–12-year-olds	215	1.80	1.76	.24	.02
Class size (R with ICALT31 = .16; sign. = .001)					
< 11	34	1.04	1.54		
11–15	40	1.44	1.80	.24	
16–20	71	1.56	1.85	.07	
21–25	125	1.53	1.52	.02	
26–30	105	1.79	1.57	.17	
> 30	25	2.58	1.91	.48	> 30 with < 11 .01

Teachers in the older student age groups show significantly better teaching skills than teachers working with younger students. The effect size difference is $\delta = .24$.

We found a small ($R = .16$) yet significant ($p = .001$) correlation between class size and teaching skill. Differences between very small class sizes (< 11 students) and very large class sizes are significant, with an effect size of $\delta = .48$. We found small effect size differences for class sizes < 11 and between 11 to 15 students and for class sizes between 21 to 25 students and classes with 26–30 students.

Conclusion and discussion

We found a reliable Rasch scale with 31 items for measuring the teaching skills of teachers working with 6 to 12-year-old students. The scale is suitable for distinguishing seven zones that give an indication of the zone of proximal development of an observed teacher.

We found no differences in the average scores of male and female teachers. As expected, teachers who just started teaching and teachers with less than 5 years of experience showed lower teaching skills compared with teachers with more experience. Teachers working with older students showed better teaching skills compared with teachers working with younger students. We found a significant though small positive correlation between class size and teaching skill. This may be related to the quality of leadership at schools, or to school policies in which the most skilled teachers are placed in larger classrooms.

For observing teaching in kindergarten, it appears that it would be better to use the 24-item version of the ICALT scale, as presented in Appendix 2.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Wim J. C. M. van de Grift (1951) is emeritus professor in Educational Sciences at the University of Groningen. He was the director of the Teacher Training Institute at the University of Groningen and scientific advisor of the Inspectorate of Education in the Netherlands. His research programme is aimed at studying the professional development of teachers in different countries.

Thoni A. M. Houtveen (1952) is emeritus professor in Literacy Research at the Utrecht University of Applied Sciences. Her research programme focusses on developing, implementing, and evaluating literacy intervention programmes regarding beginning, fluency, and comprehension reading in elementary education. Data-driven feedback is a major tool in improving the quality of instruction of the teachers involved.

Henk T. G. van den Hurk (1956), PhD, works as an educational researcher at the Utrecht University of Applied Sciences. His research projects relate to the professional development of teachers through the application of standardized observational tools in combination with data feedback on educational behaviour.

Oscar Terpstra (1982), MSc, is an educational researcher working for the Archimedes Institute at the Utrecht University of Applied Sciences. His research interest focusses on professional development of teachers in primary and secondary education with standardized observation instruments.

ORCID

Wim J. C. M. van de Grift http://orcid.org/0000-0001-9459-5292
Thoni A. M. Houtveen http://orcid.org/0000-0003-0758-2325

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*, 95–135. doi:10.1086/508733

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140. doi:10.1007/BF02291180

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69–81. doi:10.1007/BF02293746

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Brandsma, H. P., & Knuver, J. W. M. (1989). Effects of school and classroom characteristics on pupil progress in language and arithmetic. *International Journal of Educational Research*, *13*, 777–788. doi:10.1016/0883-0355(89)90028-1

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*, 245–281. doi:10.3102/00346543065003245

Cai, L., Thissen, D., & Du Toit, S. (2005–2013). IRTPRO (Version 2.1) [Computer software]. Lincolnwood, IL: Scientific Software.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics*, *24*, 323–341. doi:10.2307/1165366

Capie, W., Johnson, C. E., Anderson, S. J., Ellet, C., & Okey, J. R. (1980). *Teacher performance assessment instruments*. Athens, GA: University of Georgia.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*, 462–494. doi:10.1177/0049124108314720

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289. doi:10.3102/10769986022003265

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Commissie Evaluatie Basisonderwijs. (1994). *Inhoud en opbrengsten van het basisonderwijs* [Curriculum, process and output of elementary schools]. De Meern: Inspectie van het Onderwijs.

Cotton, K. (1995). *Effective schooling practices: A research synthesis 1995 update*. Portland, OR: Northwest Regional Educational Laboratory.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*, 215–242.

Creemers, B. P. M. (1991). *Effectieve instructie* [Effective instruction]. Den Haag: SVO.

Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.

Danielson, C. (2011). *The Framework for Teaching Evaluation Instrument: 2011 Louisiana edition*. Princeton, NJ: The Danielson Group.

Department for Education. (2012). *Teacher appraisal and capability: A model policy for schools* (DFE-57518-2012). Retrieved from www.gov.uk/government/publications

Doherty, K. M., & Jacobs, S. (2013). *State of the states 2013: Connect the dots: Using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality. Retrieved from https://www.nctq.org/dmsView/State_of_the_States_2013_Using_Teacher_Evaluations_NCTQ_Report

Ellis, E. S., & Worthington, L. A. (1994). *Research synthesis on effective teaching principles and the design of quality tools for educators* (Technical Report No. 5). Eugene, OR: University of Oregon, National Center to Improve the Tools of Educators.

Evertson, C. (1987). *Classroom activity record: Observation record for project STAR*. Nashville, TN: Vanderbilt University.

Evertson, C. M., & Burry, J. A. (1989). Capturing classroom context: The observation instrument as lens for assessment. *Journal of Personnel Evaluation in Education*, *2*, 297–320. doi:10.1007/BF00139647

Fischer, G. H., & Schleiblechner, H. H. (1970). Algorithmen und Programmen für das probabilistische Testmodel von Rasch [Algorithms and programs for the probabilistic test model of Rasch]. *Psychologische Beiträge*, *12*, 23–51.

Flanders, N. A. (1961). *Interaction analysis: A technique for quantifying teacher influence*. Minneapolis, MN: University of Minnesota, College of Education, Bureau of Educational Research.

Flanders, N. A. (1970). *Analyzing teaching behavior*. Reading, MA: Addison-Wesley.

Florida Coalition for the Development of a Performance Measurement System. (1983). *Domains: Knowledge base of the Florida performance measurement system*. Tallahassee, FL: Office of Teacher Education, Certification and In-service Staff Development.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112. doi:10.3102/003465430298487

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430–511. doi:10.1080/07370000802177235

Hill, H. C., Umland, K., Litke, E., & Kapitula, L. R. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education*, *118*, 489–519. doi:10.1086/666380

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 53–68). New York, NY: Springer.

Houtveen, A. A. M. (1990). *Begeleiden van vernieuwingen* [Supporting educational innovations]. De Lier: Academisch Boeken Centrum.

Houtveen, A. A. M., Booij, N., De Jong, R., & Van de Grift, W. J. C. M. (1999). Adaptive instruction and pupil achievement. *School Effectiveness and School Improvement*, *10*, 172–192. doi:10.1076/sesi.10.2.172.3508

Houtveen, A. A. M., & Overmars, A. M. (1996). *Instructie bij rekenen en wiskunde* [Instruction in mathematics education]. Utrecht: ISOR.

Houtveen, A. A. M., & Van de Grift, W. J. C. M. (2007a). Effects of metacognitive strategy instruction and instruction time on reading comprehension. *School Effectiveness and School Improvement*, *18*, 173–190. doi:10.1080/09243450601058717

Houtveen, A. A. M., & Van de Grift, W. J. C. M. (2007b). Reading instruction for struggling learners. *Journal of Education for Students Placed at Risk*, *12*, 405–424. doi:10.1080/10824660701762001

Houtveen, A. A. M., Van de Grift, W. J. C. M., & Brokamp, S. K. (2014). Fluent reading in special elementary education. *School Effectiveness and School Improvement*, *25*, 555–569. doi:10.1080/09243453.2013.856798

Houtveen, A. A. M., Van de Grift, W. J. C. M., & Creemers, B. P. M. (2004). Effective school improvement in mathematics. *School Effectiveness and School Improvement*, *15*, 337–376. doi:10.1080/09243450512331383242

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi:10.1080/10705519909540118

Inspectie van het Onderwijs. (1998). *Schooltoezicht primair onderwijs* [Inspection of primary education]. De Meern: Author.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: "Aufgabenkultur" und Unterrichtsgestaltung [Mathematics lessons in secondary I: "Task culture"

and teaching design]. In E. Klieme & J. Baumert (Eds.), *TIMSS – Impulse für Schule und Unterricht* (pp. 43–57). Bonn: Bundesministerium für Bildung und Forschung.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254–284. doi:10.1037/0033-2909.119.2.254

Levine, D. U., & Lezotte, L.W. (1990). *Unusually effective schools: A review and analysis of research and practice*. Madison, WI: The National Center for Effective Schools Research and Development.

Levine, D. U., & Lezotte, L. W. (1995). Effective schools research. In J. A. Banks & C. A. M. Banks (Eds.), *Handbook of research on multicultural education* (pp. 525–547). New York, NY: Macmillan.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Liu, I.-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics*, *52*, 1223–1234. doi:10.2307/2532838

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20. doi:10.18637/jss.v020.i09

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58*(303), 690–700. doi:10.1080/01621459.1963.10500879

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 391–410. doi:10.1037/0033-2909.103.3.391

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341. doi:10.1207/s15328007sem1103_2

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. J. C. M. (2015). A longitudinal study of induction on the acceleration of growth in teaching quality of beginning teachers through the eyes of their students. *Teaching and Teacher Education*, *51*, 225–245. doi:10.1016/j.tate.2015.07.003

Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better*. New York, NY: McKinsey & Company. Retrieved from https://www.mckinsey.com/industries/social-sector/our-insights/how-the-worlds-most-improved-school-systems-keep-getting-better

Muijs, D., & Reynolds, D. (Eds.). (2010). *Effective teaching: Evidence and practice* (3rd ed.). London: Sage.

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Authors.

Office for Standards in Education. (1995). *Guidance on the inspection of nursery and primary schools*. London: Author.

Organisation for Economic Co-operation and Development. (2012). *PISA 2012 results in focus: What 15-year-olds know and what they can do with what they know*. Paris: Author.

Organisation for Economic Co-operation and Development. (2014). *TALIS 2013 results: An international perspective on teaching and learning*. Paris: Author.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement*, *40*, 353–370. doi:10.1111/j.1745-3984.2003.tb01151.x

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS)*. Baltimore, MD: Brookes.

Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. *The Elementary School Journal*, *83*, 427–452. doi:10.1086/461325

Purkey, S. C., & Smith, M. S. (1985). School reform: The district policy implications of the effective schools literature. *The Elementary School Journal*, *85*, 352–389. doi:10.1086/461410

Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology*. Copenhagen: Danish Institute for Educational Research.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software*, *17*(5), 1–25. doi:10.18637/jss.v017.i05

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252. doi:10.1257/0002828041302244

Roeleveld, J. (2003). *Herkomstkenmerken en begintoets: Secundaire analyses op het PRIMA-cohort onderzoek* [Social background and testing in the early years: Secondary analyses on the PRIMA cohort study]. Amsterdam: SCO Kohnstamm Instituut.

Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. London: Office for Standards in Education.

Schaffer, E. C., & Nesselrodt, P. S. (1992, April). *The development and testing of the special strategies observation system*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Scheerens, J. (1989). *Wat maakt scholen effectief? Samenvattingen en analyses van onderzoeksresultaten* [What explains a school's effectivity? Summaries and analyses of research outcomes]. 's-Gravenhage: Instituut voor Onderzoek van het Onderwijs SVO.

Scheerens, J. (1992). *Effective schooling: Research, theory and practice*. London: Cassell.

Scheerens, J. (2008). *Een overzichtsstudie naar school- en instructie-effectiviteit: Samenvattingen en analyses van onderzoeksresultaten* [Review of school and instruction effectiveness: Summaries and analyses of research outcomes]. Enschede: Universiteit Twente.

Slavin, R. E. (1987). Ability grouping and achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, *57*, 293–336. doi:10.3102/00346543057003293

Stallings, J. (1980). Allocated academic learning time revisited, or beyond time on task. *Educational Researcher*, *9*(11), 11–16. doi:10.3102/0013189X009011011

Stallings, J. A., & Kaskowitz, D. H. (1974). *Follow through classroom observation evaluation, 1972–1973*. Menlo Park, CA: SRI International.

Teddlie, C., Creemers, B., Kyriakides, L., Muijs, D., & Yu, F. (2006). The international system for teacher observation and feedback: Evolution of an international study of teacher effectiveness constructs. *Educational Research and Evaluation*, *12*, 561–582. doi:10.1080/13803610600874067

Teddlie, C., Virgilio, I., & Oescher, J. (1990). Development and validation of the Virgilio teachers' behavior instrument. *Educational and Psychological Measurement*, *50*, 421–430. doi:10.1177/0013164490502021

Thurlings, M. C. G. (2012). *Peer to peer feedback: A study on teachers' feedback processes*. Maastricht: Universitaire Pers Maastricht.

Tricket, E. J., & Moos, R. H. (1974). *The Classroom Environment Scale (CES)*. Palo Alto, CA: Consulting Psychologists Press.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. doi:10.1007/BF02291170

Van de Grift, W. (1985). Onderwijsleerklimaat en leerlingprestaties [Educational climate and student achievement]. *Pedagogische Studiën*, *62*, 401–414.

Van de Grift, W. J. C. M. (1994). *Technisch rapport van het onderzoek onder 386 basisscholen ten behoeve van de evaluatie van het basisonderwijs* [Report on the evaluation of 386 schools for primary education]. De Meern: Inspectie van het Onderwijs.

Van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and an application of an assessment instrument. *Educational Research*, *49*, 127–152. doi:10.1080/00131880701369651

Van de Grift, W. J. C. M. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement*, *25*, 295–311. doi:10.1080/09243453.2013.794845

Van de Grift, W., & Helms-Lorenz, M. (2013). Waarom verlaten zoveel beginnende leraren de school waar ze hun carrière begonnen? [Why do so many beginning teachers leave the school where they started their careers?]. *Van Twaalf tot Achttien*, *23*(8), 12–15.

Van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation*, *43*, 150–159. doi:10.1016/j.stueduc.2014.09.003

Van de Grift, W. J. C. M., & Lam, J. F. (1998). Het didactisch handelen in het basisonderwijs [Teaching in primary education]. *Tijdschrift voor Onderwijsresearch*, *23*(3), 224–241.

Van de Grift, W., Van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogisch didactische vaardigheid van leraren in het basisonderwijs [Development of teaching skills in primary education]. *Pedagogische Studiën*, *88*, 416–432.

Van den Hurk, H. T. G., Houtveen, A. A. M., & Van de Grift, W. J. C. M. (2016). Fostering effective teaching behaviour through the use of data-feedback. *Teaching and Teacher Education*, *60*, 444–451. doi:10.1016/j.tate.2016.07.003

Van der Lans, R. M., Van de Grift, W. J. C. M., & Van Veen, K. (2018). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education*, *86*, 247–264. doi:10.1080/00220973.2016.1268086

Van der Lans, R. M., Van de Grift, W. J. C. M., Van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, *50*, 88–95. doi:10.1016/j.stueduc.2016.08.001

Veenman, S., Lem, P., Voeten, B., Winkelmolen, B., & Lassche, H. (1986). *Onderwijs in combinatieklassen* [Education in multigraded classrooms]. 's-Gravenhage: SVO.

Virgilio, I. (1987). *An examination of the relationships among school effectiveness in elementary and junior high schools* (Doctoral dissertation). New Orleans, LA: University of New Orleans.

Virgilio, I., & Teddlie, C. (1989). *Technical manual for the Virgilio Teacher Behavior Inventory* (Unpublished manuscript). University of New Orleans, New Orleans, LA.

von Davier, M. (1994). *WINMIRA: A program system for analyses with the Rasch model, with the latent class analysis and with the mixed Rasch model*. Kiel: Institute for Science Education (IPN).

Vygotskij, L. S. (2002). *Denken und Sprechen* [Thinking and speech]. Weinheim: Beltz Verlag. (Original work published 1934)

Walberg, H. J., & Haertel, G. D. (1992). Educational psychology's first century. *Journal of Educational Psychology*, *84*, 6–19. doi:10.1037/0022-0663.84.1.6

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika*, *54*, 427–450. doi:10.1007/BF02294627

Wijnstra, J., Ouwens, M., & Béguin, A. (2003). *De toegevoegde waarde van de basisschool* [Added value of schools in elementary education]. Arnhem: CITOgroep.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*, 57–67. doi:10.1023/A:1007999204543

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, *10*, 321–344. doi:10.1207/s15324818ame1004_2

Appendix 1. International Comparative Analysis of Learning and Teaching ICALT-LESSON-OBSERVATION FORM for primary education

Country	…	Date observation (dd-mm-yyyy)	…
School name	…	Class	…
Location	…	No of learners present	…
Level of education	0 = Kindergarten 1 = elementary education	Name observer	…
School denomination	0 = public 1 = private	Gender observer	F/M
Subject matter	…	Years of teaching experience observer	…
Name teacher	…	Occupation observer	0 = school teacher 1 = university teacher 2 = other, please specify…
Gender teacher	M/F		
Years of teaching experience teacher	…		

Observe the following behaviours and events
Rate[1] Please circle the appropriate answer:
1 = mostly weak; 2 = more often weak than strong; 3 = more often strong than weak; 4 = mostly strong
Observed[2] Please circle the appropriate answer:
0 = no, I have not observed this; 1 = yes, I have observed this

Domain		Indicator: The teacher…	Rate[1]	Examples of good practice: *The teacher …*	Observed[2]
Safe and stimulating learning climate	1	**…shows respect for learners in his/her behaviour and language**	**1 2 3 4**	*…lets learners finish their sentences*	*0 1*
				…listens to what learners have to say	*0 1*
				…does not make role stereotyping remarks	*0 1*
	2	**…maintains a relaxed atmosphere**	**1 2 3 4**	*…addresses learners in a positive manner*	*0 1*
				…uses and stimulates humour	*0 1*
				…accepts the fact that learners make mistakes	*0 1*
				…shows compassion and empathy for all learners present	*0 1*
	3	**…promotes learners' self-confidence**	**1 2 3 4**	*…gives positive feedback on questions and remarks from leaners*	*0 1*
				…compliments learners on their work	*0 1*
				…acknowledges the contributions that learners make	*0 1*
	4	**…fosters mutual respect**	**1 2 3 4**	*…stimulates learners to listen to each other*	*0 1*
				…intervenes when learners make fun of someone	*0 1*
				…keeps (cultural) differences and idiosyncrasies in mind	*0 1*
				…stimulates solidarity between learners	*0 1*
				…encourages learners to experience activities as group events	*0 1*
Efficient organisation	5	**…ensures the lesson proceeds in an orderly manner**	**1 2 3 4**	*Learners enter and settle in an orderly manner*	*0 1*
				…intervenes timely and appropriately in case of disorder	*0 1*
				…safeguards the agreed rules and codes of conduct	*0 1*
				…keeps all leaners involved in activities until the end of the lesson	*0 1*
				…makes sure that learners know what to do if they need help with their work and explains clearly when they can ask for help	*0 1*
				…makes sure learners know what to do when they have finished their work	*0 1*
	6	**…monitors to ensure learners carry out activities in the appropriate manner**	**1 2 3 4**	*…checks whether learners have understood what they have to do*	*0 1*
				…provides feedback on learners' social functioning whilst carrying out a task	*0 1*
	7	**…provides effective classroom management**	**1 2 3 4**	*…explains clearly which materials can be used*	*0 1*
				The materials for the lesson are ready for use	*0 1*
				Materials are geared at the right level and developmental stage of the learners	*0 1*
	8	**…uses the time for learning efficiently**	**1 2 3 4**	*… starts the lesson on time*	*0 1*
				… does not waste time at the beginning, during, or at the end of the lesson	*0 1*
				…prevents any unnecessary breaks from occurring	*0 1*
				…does not keep learners waiting	*0 1*

(*Continued*)

(Continued).

Domain		Indicator: The teacher…	Rate[1]	Examples of good practice: *The teacher …*	Observed[2]
	9	**…engages all learners in the lesson**	**1 2 3 4**	*…creates learners assignments which stimulate active participation*	*0 1*
				…asks questions which stimulate learners to reflect	*0 1*
				…makes sure that learners listen and/or continue working	*0 1*
				…allows for 'thinking time' after asking a question	*0 1*
				…also invites learners to participate who do not volunteer to do so	*0 1*
	10	**…encourages learners to do their best**	**1 2 3 4**	*…praises learners who do their best*	*0 1*
				…makes clear that all learners should do their best	*0 1*
				…expresses positive expectations about what learners are going to achieve	*0 1*
Clear and structured instructions	11	**…presents and explains the subject material in a clear manner**	**1 2 3 4**	*…activates prior knowledge of learners*	*0 1*
				…gives staged instructions	*0 1*
				…poses questions which learners can understand	*0 1*
				…summarises the subject material from time to time	*0 1*
	12	**…gives feedback to learners**	**1 2 3 4**	*…makes clear **whether** an answer is right or wrong*	*0 1*
				*…makes clear **why** an answer is right or wrong*	*0 1*
				…gives feedback on the way in which learners have arrived at their answer	*0 1*
	13	**…during the presentation stage, checks whether learners have understood the subject material**	**1 2 3 4**	*…ask questions which stimulate learners to reflect*	*0 1*
				…checks regularly whether learners understand what the lesson is about	*0 1*
	14	**…teaches in a well-structured manner**	**1 2 3 4**	*The lesson is built up in terms of clear stages and transitions between stages*	*0 1*
				The lesson builds up logically, going from the simple to the complex	*0 1*
				Activities and assignments are connected to the materials presented during the presentation stage	*0 1*
				The lesson offers a good variety of presentation, instruction, controlled practice, free practice, and so forth.	*0 1*
	15	**…gives a clear explanation of how to use didactic aids and how to carry out assignments**	**1 2 3 4**	*…makes sure that all learners know what to do*	*0 1*
				…explains how lesson aims and assignments relate to each other	*0 1*
				…explains clearly which materials and sources can be used	*0 1*

(*Continued*)

(Continued).

Domain		Indicator: The teacher…	Rate[1]	Examples of good practice: *The teacher …*	Observed[2]
Intensive and activating teaching	16	**…offers activities and work forms that stimulate learners to take an active approach**	1 2 3 4	*…uses diverse forms of conversation and discussion*	0 1
				…offers controlled (pre-)practice	0 1
				…lets learners work in group	0 1
				…uses information and communication technology (ICT, e.g., digiboard, beamer)	0 1
				…employs a variety of instruction strategies	0 1
				…varies assignments	0 1
				…varies lesson material	0 1
				…uses materials and examples from daily life	0 1
				…asks a range of questions	0 1
	17	**…stimulates the building of self-confidence in weaker leaners**	1 2 3 4	*…gives positive feedback on questions from weaker learners*	0 1
				…displays positive expectations about what weaker learners have to achieve	0 1
				…compliments weaker learners on their works	0 1
				…acknowledges the contributions made by weaker learners	0 1
	18	**…stimulates learners to think about solutions**	1 2 3 4	*…shows learners the path they can take towards a solution*	0 1
				…teaches strategies for problem-solving and referencing	0 1
				…teaches learners how to consult sources and reference works	0 1
				…offers learners checklists for problem solving	0 1
	19	**…asks questions which stimulate learners to reflect**	1 2 3 4	*…waits long enough to give all learners the chance to answer a question*	0 1
				…encourages learners to ask each other questions and explain things to each other	0 1
				…asks learners to explain the different steps of their strategy	0 1
				…checks regularly whether instructions have been understood	0 1
				…asks questions which stimulate reflection and learner feedback	0 1
				…checks regularly whether learners understand what the lesson is about	0 1
	20	**…lets learners think aloud**	1 2 3 4	*…provides the opportunity for learners to think aloud about solutions*	0 1
				…asks learners to verbalise solutions	0 1
	21	**…gives interactive instructions**	1 2 3 4	*…promotes the interaction between learners*	0 1
				…promotes the interaction between teacher and learners	0 1
	22	**…clearly specifies the lesson aims at the start of the lesson**	1 2 3 4	*…informs learners at the start of the lesson about the lesson aim*	0 1
				…clarifies the aims of assignments and their learning purpose	0 1
Adjusting instructions and learner processing to inter-learner differences	23	**…evaluates whether the lesson aims have been reached**	1 2 3 4	*…evaluates whether the lesson aims have been reached*	0 1
				…evaluates learners' performance	0 1

(*Continued*)

(Continued).

Domain		Indicator: The teacher…	Rate[1]	Examples of good practice: *The teacher …*	Observed[2]
	24	**…offers weaker learners extra study and instruction time**	**1 2 3 4**	*…gives weaker learners extra study time*	*0 1*
				…gives weaker learners extra instruction time	*0 1*
				…gives weaker learners extra exercises/practices	*0 1*
				…gives weaker learners "pre- or post-instruction"	*0 1*
	25	**…adjusts instructions to relevant inter-learner differences**	**1 2 3 4**	*…puts learners who need little instructions (already) to work*	*0 1*
				…gives additional instructions to small groups or individual learners	*0 1*
				…does not simply focus on the average learner	*0 1*
	26	**…adjusts the processing of subject matter to relevant inter-learner differences**	**1 2 3 4**	*…distinguishes between learners in terms of the length and size of assignments*	*0 1*
				…allows for flexibility in the time learners get to complete assignments	*0 1*
				…lets some learners use additional aids and means	*0 1*
Teaching learning strategies	**27**	**…teaches learners how to simplify complex problems**	**1 2 3 4**	*…teaches learners how to simplify complex problem*	*0 1*
				…teaches learners how to break down complex problems into simpler ones	*0 1*
				…teaches learners to order complex problem	*0 1*
	28	**…stimulates the use of control activities**	**1 2 3 4**	*…pays attention to prediction strategies for reading*	*0 1*
				…lets learners relate solutions to the context of a problem	*0 1*
				…stimulates the application of alternative strategies	*0 1*
	29	**…teaches learners to check solutions**	**1 2 3 4**	*…teaches learners how to estimate outcomes*	*0 1*
				…teaches learners how to predict outcomes	*0 1*
				…teaches learners how to relate outcomes to the practical context	*0 1*
	30	**…stimulates the application of what has been learned**	**1 2 3 4**	*… stimulates the conscious application of what has been learned in other (different) learning contexts*	*0 1*
				…explains to learners how solutions can be applied in different situations	*0 1*
				…relates problems to previously solved problem	*0 1*
	31	**…encourages learners to think critically**	**1 2 3 4**	*…asks learners to provide explanations for occurrences*	*0 1*
				…asks learners for their opinion	*0 1*
				…asks learners to reflect on solutions or answers given	*0 1*
				…asks learners to provide examples of their own	*0 1*
	32	**…asks learners to reflect on practical strategies**	**1 2 3 4**	*…asks learners to explain the different steps of the strategy applied*	*0 1*
				…gives an explicit explanation of possible (problem-solving) strategies	*0 1*

(*Continued*)

(Continued).

Domain		Indicator: The teacher…	Rate[1]	Examples of good practice: *The teacher …*	Observed[2]
				…asks learners to expand on the pros and cons of different strategies	*0 1*
Learner engagement	**33**	**…are fully engaged in the lesson**	**1 2 3 4**	*…pay attention when instructions are given*	*0 1*
				…participate actively in conversations and discussions	*0 1*
				…ask questions	*0 1*
	34	**…show that they are interested**	**1 2 3 4**	*…listen actively when instructions are being given*	*0 1*
				…show their interest by asking follow-up questions	*0 1*
	35	**…take an active approach to learning**	**1 2 3 4**	*…ask follow-up questions*	*0 1*
				…show that they take responsibility for their own learning process	*0 1*
				…work independently	*0 1*
				…take the initiative themselves	*0 1*
				…use their time efficiently	*0 1*

Version EE2018
Any reactions or comments? Please contact: Wim.vandeGrift@ZIGGO.NL

Appendix 2. The ICALT24 version for use in Kindergarten and elementary education

The eRm R package (Mair & Hatzinger, 2007) is used to compute the difficulty parameters for each of the 24 selected items. The program WINMIRA (von Davier, 1994) is used to compute the person parameters Warm's weighted likelihood estimates. Table A1 shows the Wright map of the ICALT scale with the difficulty parameters and the standard errors of the items and the person parameters and their standard errors and relative frequency.

Table A1. Wright map for the ICALT24 scale (N = 500 teachers teaching 4–12-year-olds).

		Item parameters		Person parameters		
		Log δ	*SE*	Warm's θ	*SE*	%
Beyond the basic standards of teaching				−5.28	1.60	.0
				−3.96	0.97	.0
	Shows respect for learners in his/her behaviour and language	−3.61	.42			
				−3.25	0.79	.0
				−2.73	0.69	.2
	Maintains a relaxed atmosphere	−2.70	.29			
				−2.31	0.63	.2
				−1.95	0.59	.8
	Promotes learners' self-confidence	−1.85	.21			
	Provides effective classroom management	−1.77	.20			
	Presents and explains the subject material in a clear manner	−1.73	.20			
				−1.63	0.56	.4
	Teaches in a well-structured manner	−1.58	.19			
				−1.34	0.54	1.4
	Ensures the lesson proceeds in an orderly manner	−1.32	.18			
				−1.06	0.53	1.2
				−0.79	0.52	1.0
	Uses the time for learning efficiently	−.73	.15			
Basic teaching skills	Fosters mutual respect	−.58	.15			
				−0.53	0.51	3.6
	Engages all learners in the lesson	−.37	.14			
	Encourages learners to do their best	−.31	.14			
				−0.28	0.51	2.4
	Monitors to ensure learners carry out activities in the appropriate manner	−.10	.13			
	Gives a clear explanation of how to use didactic aids and how to carry out assignments	−.19	.13	−0.02	0.51	4.2
				0.23	0.51	4.0
	Gives feedback to learners	.18	.13			

(*Continued*)

Table A1. (Continued).

		Item parameters		Person parameters		
		Log δ	*SE*	Warm's θ	*SE*	%
Activating students	Offers activities and work forms that stimulate learners to take an active approach	.44	.12	0.49	0.52	8.2
	Asks questions which stimulate learners to reflect	.54	.12			
	Gives interactive instructions	.58	.12			
				0.76	0.53	4.6
	Let learners think aloud	.84	.12	1.04	0.54	6.2
				1.34	0.56	7.4
Differentiating teaching	Evaluates whether the lesson aims have been reached	1.63	.11	1.66	0.58	6.6
				1.99	0.60	7.4
	Stimulates the application of what has been learned	2.09	.11			
	Adjusts the processing of subject matter to relevant inter-learner differences	2.33	.12			
				2.37	0.64	9.4
teaching learning strategies	Encourages learners to think critically	2.50	.12			
				2.79	0.69	8.8
	Stimulates the use of control activities	2.82	.12			
	Teaches learners to check solutions	2.85	.12			
Almost perfect				3.28	0.77	7.8
				3.94	0.94	5.8
Perfect				5.18	1.55	8.4

Table A2 shows the average scores on the ICALT24 scale of the total sample of 500 teachers working in 4–5-year-olds, 5–7-year-olds, and 8–12-year-olds.

Table A2. Descriptives of the ICALT24 scale.

	Sample (N = 500)		Kindergarten (n = 96)		6–7-year-olds (n = 184)		8–12-year-olds (n = 215)	
	μ	σ	μ	σ	μ	σ	μ	σ
Measurement 1	1.57	1.66	1.41	1.73	1.39	1.52	1.80	1.72
Cohen's δ					−.01		.25	
Significance					n.s.		.01	

Note: Due to missing values, the addition of Kindergarten, 6–7-year-olds and 8–12-year-olds is lower than 500.